Volume Title: The Measurement of Labor Cost

Volume Author/Editor: Jack E. Triplett, ed.

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-81256-1

Volume URL: http://www.nber.org/books/trip83-1

Publication Date: 1983

Chapter Title: Imputing Income in the CPS: Comments on â€œMeasures of Aggregate Labor Cost in the United Statesâ€•

Chapter Author: Donald Rubin

Chapter URL: http://www.nber.org/chapters/c7382

Chapter pages in book: (p. 333 - 344)

9 Imputing Income in the CPS: Comments on "Measures of Aggregate Labor Cost in the United States"

Donald B. Rubin

9.1 Introduction

Three of the four papers in this section (Gollop and Jorgenson, Smeeding, and Johnson) base conclusions on income data from the Census Bureau's Current Population Survey (CPS). The CPS is a major source of income data for economic research even though the nonresponse rate on income items is about 15–20%. This level of nonreporting of income, especially if concentrated among special types of individuals, should be of substantial concern to researchers in economics. Most published economic research, however, ignores this problem when using CPS data. The major reason that researchers can ignore this nonreporting of income is that before CPS public-use tapes are released, the Census Bureau imputes (i.e., fills in) missing income data (as well as other data). Although imputed data are flagged to distinguish them from real data, it is evidently easy for researchers to be seduced into ignoring this distinction and treating all values, imputed and real, on the same basis.

Three recent articles on income imputation in the CPS address the adequacy of current imputation procedures. They are Lillard, Smith, and Welch (1982, hereafter LSW), Greenlees, Reece, and Zieschang (1982, hereafter GRZ), and Herzog and Rubin (1983, hereafter HR). My comments here are designed to highlight relevant issues arising from the existence of income nonreporters in the CPS, especially in the context of work presented in these articles and other recent literature.

Donald B. Rubin is professor, Department of Statistics and Department of Education, University of Chicago.

This research was sponsored by the United States Army under contract DAAG29-80-C-0041 while the author was visiting professor with the Mathematics Research Center at the University of Wisconsin, Madison.

After characterizing income nonreporters in section 9.2 and describing the Census Bureau's hot deck procedure in section 9.3, in section 9.4 I point out the need for multiple imputation if uncertainty due to nonresponse is to be properly reflected in an imputed data set. Section 9.5 provides definitions of ignorable and nonignorable nonresponse, while section 9.6 describes the selection model used in LSW and GRZ and emphasizes that external information is needed to justify the acceptance of such a model or any other particular model for nonresponse as an accurate reflection of reality. Finally, section 9.7 briefly describes the CPS-SSA-IRS Exact Match File, which both GRZ and HR use to help provide such external information.

9.2 Who Are the Nonrespondents on Income Questions?

Of central importance for determining whether the 15–20% nonresponse rate on income questions is of major concern is the extent to which income nonreporters are different from income reporters. If the nonreporters were just a simple random sample from the population of reporters and nonreporters, the loss in efficiency of estimation created by ignoring the nonreporters altogether would be of little concern.

There is a great deal of evidence, however, that nonreporters do differ from reporters in important ways. One such piece of evidence that LSW presents is especially interesting. Apparently, if we were to plot "probability of nonresponse on income items" versus "amount of actual income," the relationship would be U-shaped: moderate nonresponse at low incomes, low nonresponse at moderate incomes, and very high nonresponse at high incomes. Moreover, LSW's evidence suggests that this U-shaped relationship is created by the existence of two primary types of income nonreporters. The first type is called "general nonreporters" because they have a high nonresponse rate on many CPS questions, not just income questions. These people tend to have low incomes and approach CPS questions in a generally reluctant manner. The second type of income nonreporter is called "specific nonreporters" because on most CPS questions, that is nonincome questions, they have low nonresponse rates, whereas on income questions their nonresponse rates are very high (e.g., over 30%). The specific nonreporters tend to be professionals with high incomes, for example, doctors, lawyers, and dentists.

If we accept this interesting picture as relatively accurate, it seems to me natural and desirable to try to build a nonresponse model that explicitly recognizes the U-shaped relationship and the two types of income nonreporters. The LSW and GRZ selection models, however, do not exploit this structure and instead use models for nonresponse asserting that, conditional on some predictor variables (such as years of education), the relationship between probability of nonresponse on income

items and income is monotonic. Of course, one can criticize virtually any analysis for not fully exploiting some interesting features found in subsequent analyses. Consequently, my comment on this point should be viewed more as offering a suggestion for further study than as criticizing the work presented in LSW and GRZ.

9.3 The Census Bureau's Hot Deck Imputation Scheme

The Census Bureau's procedure for imputation, the hot deck, has been used since the early 1960s. The hot deck is a matching algorithm in the sense that for each nonrespondent, a respondent is found who matches the nonrespondent on variables that are measured for both. The variables used for the matching are all categorical, with varying numbers of levels (e.g., "gender" has two levels, "region of country" has four levels). If a match is not found, categories are collapsed and variables are deleted so that coarser matches are allowed. Eventually, every nonrespondent finds a match; the matching respondent is often called (by hot deck aficionados) "the donor" because the donor's record of values is donated to the nonrespondent to fill in all missing values in the nonrespondent's record.

The number of variables used for matching and their level of detail has expanded over the years, and imputed income can be sensitive to such rule changes. For example, between 1975 and 1976, years of education was added to the list of matching variables, and as a consequence, the imputed incomes of nonrespondents with many years of education increased substantially from 1975 to 1976. Such changes can create problems when comparing income data in different periods of time. A related problem is that even though the ideal match that is possible under the hot deck is closer now than it was years ago, many nonrespondents fail to find donors at this ideal level of detail. For one example, only 20% find donors in the same region of the country. For a second example, judges with ideal matches are imputed to earn approximately $30,000 more than judges without ideal matches.

The hot deck, by trying for exact multivariate categorical matches, is trying to control all higher order interactions among the matching variables. This task is very difficult with many matching variables when using a categorical matching rule, even if there is a large pool of potential matches for nonrespondents. For example, suppose all the matching variables are dichotomous and independent, with 50% of the population at each level of each variable. If p is the number of matching variables $(.5)^p$ is the probability of two randomly chosen units matching each other on all p variables. Since $.5^{10} < .001$, it is obvious that finding exact matches with many matching variables, even in this ideal setting, requires very large pools of potential matches.

Related work on matching methods in observational studies investi-

gates categorical matching methods and offers alternative matching methods (e.g., Cochran and Rubin 1973; Rubin 1976*a*, 1976*b*, 1980*a*; Rosenbaum and Rubin 1983). I suspect that some of the more recent work (e.g., Rosenbaum and Rubin 1983) may have useful suggestions for an improved hot-deck-like procedure. Neither LSW nor GRZ suggests modifying the matching algorithm but rather suggests using explicit statistical models for imputation. HR considers both explicit models and hot deck procedures.

9.4 Imputation and the Need for Multiple Imputation

LSW and GRZ both suggest a model-based alternative to hot deck imputation: (a) build an explicit model, specifically, a selection model (cf. Heckman 1979) where the probability of nonresponse on income increases with income (see section 9.6 for details), (b) estimate the parameters of this model by maximum likelihood, and (c) impute one value for each missing value by randomly drawing observations from this model with unknown parameters replaced by their maximum likelihood estimates.

I have several general comments to make on imputation whether based on implicit models like the hot deck or explicit models like the selection model.

First, for the data producer, some form of imputation is almost required and often desirable even if not required. I believe the Census Bureau feels it cannot produce public-use files with blanks. Also, I believe it feels, and rightly so, that it knows more about the missing data than the typical user of public-use files. Furthermore, the typical user of public-use files will not have the statistical sophistication needed to routinely apply model-based methods for handling nonresponse, such as those reveiwed by Little (1982). Of course, in any public-use file, all imputed values must be flagged to distinguish them from real values.

Second, imputation based on explicit modeling efforts may require much more work than implicit models, such as the hot deck (or some other matching method for imputation), that can impute all missing variables at once no matter what the pattern of missing variables. Of course, this does not mean that explicit models should be avoided: explicit model-based methods are, in principle, the proper ones to handle nonresponse.

Third, when drawing values to impute, in order to obtain inferences with the correct variability, parameters of models must not be fixed at estimated values but must be drawn in such a way as to reflect uncertainty in their estimation.

Fourth, one imputation for each missing value, even if drawn according to the absolutely correct model, will lead to inferences that underestimate variability (e.g., underestimate standard errors).

Fifth, there exists a need to display sensitivity of answers to plausible models for the process that creates nonresponse since the observed data alone cannot determine which of a variety of models is correct.

These points are all leading to the suggestion to use multiple imputation as proposed in Rubin (1978*a*) and expanded upon in Rubin (1980*b*). Whether using an implicit model, such as the hot deck, or an explicit model, such as employed in LSW and GRZ, if imputation is used to handle nonresponse, multiple imputation is generally needed to reach the correct inference.

Multiple imputation replaces each missing value by a pointer to a vector, say of length m, of possible values; the m values reflect uncertainty for the correct value. Imputing only one value can only be correct when there is no uncertainty, but if there were no uncertainty, the missing value would not be missing; consequently, multiple imputation rather than single imputation is needed when there are missing data.

The m possible values for each of the missing data result in m complete data sets, and these can be analyzed by standard complete-data methods to arrive at valid inferences. Suppose, for example, that the m imputations were all made under one model for nonresponse, such as the LSW selection model, and suppose that with complete data we would form the estimate $\hat{Q}$ with associated standard error S. Let $\hat{Q}_i$ and S_i, $i = 1, \ldots, m$, be their values in each of the data sets created by miltiple imputation. Then the resultant multiple imputation estimate is simply $\bar{Q} = \Sigma\hat{Q}_i/m$ with standard error

$$\sqrt{\Sigma(\hat{Q}_i - \bar{Q})^2/(m-1) + \Sigma S_i^2/m}.$$

If the m imputations are from k different models, then those imputations under each model should be combined to form one inference under each model, and then the comparison across the k resulting inferences displays sensitivity of inference to the k different models.

HR applies multiple imputation to the CPS and compares the results with single imputation answers. Both an explicit model and a hot deck procedure are considered. In contrast to both LSW and GRZ, the income variable being imputed in HR is not total income, but rather social security benefits. Also, the model used in HR is not a selection model, but rather a two-stage log-liner/linear model, where the log-linear model is used to predict the existence of social security benefits (a 0-1 variable), and the linear model is used to predict the amount of benefits (actually, log benefits), given that some benefits were received. This work illustrates that multiple imputation can play an important practical role.

9.5 The Distinction between Ignorable Nonresponse and Nonignorable Nonresponse

An important distinction between the LSW and GRZ selection models and the HR two-stage model involves underlying assumptions. Models

for survey nonresponse can be classified into ones assuming "ignorable" nonresponse and those assuming "nonignorable" nonresponse, the terminology being from Rubin (1976*c*, 1978*b*). I believe that LSW's use of "random nonresponse" is intended to convey essentially the same notion, although I find the LSW use of this phrase somewhat inconsistent. Both GRZ and HR use the ignorable/nonignorable classification.

Under ignorable nonresponse models, respondents and nonrespondents that are exactly matched with respect to observed variables have the same distribution of missing variables. The Census Bureau hot deck operates under this assumption, although it does not have to do so. For example, having found a donor for a nonrespondent, instead of imputing the donor's income, the hot deck algorithm could be instructed to impute the donor's income *plus* 10 percent. If we accept the Census Bureau's hot deck as currently implemented, then we implicitly accept the hypothesis that nonresponse is ignorable, and then there is no need to be concerned with selection models, such as used in LSW and GRZ. Instead, under ignorable nonresponse, all energy should be focused on modeling the conditional distribution of missing variables given observed variables for respondents, since, by assumption, this conditional distribution is the same for nonrespondents and respondents. The explicit model in HR posits ignorable nonresponse and focuses on predicting, for respondents, the amount of social security benefits.

When missing values are to be replaced by imputed values, however, whether these values arise from implicit or explicit models, a single imputation generally will underestimate variability. Consequently, the LSW statement accepting the hot deck if operating at its most detailed level is not entirely appropriate if valid inferences are desired, even if nonresponse is ignorable. Both GRZ and HR explicitly acknowledge this point, and HR uses multiple imputation under ignorable nonresponse models to address it.

Under nonignorable nonresponse models, respondents and nonrespondents perfectly matched on observed variables have different distributions on unobserved variables. The example of the modified hot deck which imputes donor's income plus 10 percent is an implicit nonignorable nonresponse model; the LSW and GRZ selection models are explicit nonignorable models since the probability of nonresponse increases with income. When nonignorable nonresponse is possible, as with income nonreporting in the CPS, it is crucial to expose sensitivity of answers to different models, all of which are consistent with the data. An important contribution of LSW is that it defines and illustrates the use of an expanded collection of such models. Specifically, LSW extends the GRZ selection model in which log(total income) is normally distributed to a selection model in which some Box-Cox (1964) transformation of total income is normally distributed, where the transformation is to be estimated.

Within the context of imputation for missing values, sensitivity to models can only be exposed through the use of multiple imputation, where for each missing value there are imputations under each model being considered (e.g., two imputations under the ignorable hot deck, two imputations under the nonignorable [plus 10 percent] model, and two imputations under the GRZ nonignorable selection model). Again, such multiple imputations are necessary to reach valid inferences under each model and to expose sensitivity of answers to population features not addressable by the observed data.

9.6 An Explicit Nonignorable Model: Caveats and Results

Let Y be earnings, which is sometimes missing in the CPS, and let X be a vector of predictor variables (e.g., education, work experience), which for simplicity is assumed to be always observed in the CPS. Define Y^* to be the Box-Cox (1964) transformed earnings ($Y^* = [Y^\theta - 1]/\theta$), Z to be an unobserved, hypothetical variable such that Y is missing if $Z > 0$, and suppose (Y^*, Z) given X is bivariate normal with correlation ρ. Since CPS income is "top coded" at \$50,000, if Y is greater than \$50,000 and $Z \leq 0$, then the observed income is \$50,000. The parameters of the regression of (Y^*, Z) on X as well as θ and ρ are unknown and to be estimated.

If $\rho = 0$, nonresponse is ignorable, whereas if $\rho \neq 0$, nonresponse is nonignorable; as $|\rho| \to 1$, the extent of nonignorable nonresponses becomes more serious in the sense that the distribution of Y^* residuals for respondents becomes less normal and more skewed. This defines the LSW model, and LSW obtains maximum likelihood estimates for all parameters, explicitly recognizing the truncation of Y at \$50,000 in the CPS. Essentially the same model with the restriction that $\theta = 0 (Y^* = \log[Y])$ is used in GRZ. The extension to other θ is certainly interesting and potentially quite useful. Of particular importance, it gives users a broader range of models for nonresponse to which sensitivity of estimation can be investigated.

It must not be forgotten, however, that the estimation of parameters is relying critically on the assumed normality of the regression of (Y^*, Z) on X: both θ and ρ are chosen by maximum likelihood to make the residuals in this regression look as normal as possible. If in the real world there is no (θ, ρ) that makes this regression like a normal linear regression, then there is no real reason to believe that the answers that are obtained by maximizing θ and ρ lead to better real world answers. A small artificial example I've used before (Rubin 1978*a*) illustrates this point in a simpler context:

> Suppose that we have a population of 1000 units, try to record a variable Z, but half of the units are nonrespondents. For the 500 respondents, the data look half-normal. Our objective is to know the

> mean of Z for all 1000 units. Now, if we believe that the nonrespondents are just like the respondents except for a completely random mechanism that deleted values (i.e., if we believe that mechanisms are ignorable), the mean of the respondents, that is, the mean of the half-normal distribution, is a plausible estimate of the mean for the 1000 units of the population. However, if we believe that the distribution of Z for the 1000 units in the population should look more or less normal, then a more reasonable estimate of the mean for the 1000 units would be the minimum observed value because units with Z values less than the mean refused to respond. Clearly, the data we have observed cannot distinguish between these two models except when coupled with prior assumptions. (p. 22)

Notwithstanding the above caveats, suppose we put our faith in the normal linear model for the bivariate regression of (Y^*, Z) on X. LSW produces some interesting results using white males, 16–65 years old, in the 1970, 1975, 1976, and 1980 CPS. One interesting, but not surprising, result is that fixing θ at 1 ($Y^* = Y$) produces very different answers from fixing θ at 0 ($Y^* = \log[Y]$); if $\theta = 1$, nonrespondents are imputed to earn less than matching respondents, whereas if $\theta = 0$, nonrespondents are imputed to earn more than matching respondents. With θ fixed, the asymmetry in the Y^* given X residuals addresses the correlation ρ and so determines the extent to which the nonresponse is nonignorable. Thus, we have learned that the Y given X residuals are skewed left and the $\log(Y)$ given X residuals are skewed right. Further study shows that $\theta = .45$ provides a better fit to the data than either $\theta = 0$ or $\theta = 1$, but that the residuals are still skewed right; under $\theta = .45$ we find that nonrespondents are imputed to earn more than similar respondents; $\theta = .45$ leads to a 10% increase in average earnings over the CPS hot deck values, \$18,000 versus \$16,000.

But we must remember that if the distribution of $Y^{(.45)}$ given X really has the right asymmetry that is observed when $Y^{(.45)}$ is regressed on X, then the adjustment created by assuming a selection effect on Z is entirely inappropriate, and (just as with the artificial half-normal example) the data cannot distinguish between the ignorable and nonignorable alternatives. More precisely, suppose first that, in the population, $Y^{(.45)}$ has a linear regression on X with a skew distribution of residuals like that observed when we regress $Y^{(.45)}$ on X for the CPS data and that nonresponse is ignorable; such a model would generate data just like those we have observed, and then we should *not* be imputing higher incomes for nonrespondents than respondents with the same X values.

In contrast, suppose that $Y^{(.45)}$ in the population really has a normal linear regression on X and that the stochastic censoring implied by the LSW nonresponse model is correct, that is, nonresponse is nonignorable with this particular form; then, as LSW shows, we should be imputing higher incomes for nonrespondents than respondents with the same X

values. There is no way that the observed data can distinguish between these two alternatives; if we really believe Y^* given X in the population is *normal* for some θ, then we can correctly assert that the CPS hot deck procedure is biased. If we admit the possibility that Y^* given X is not normal or even symmetric for any θ, then we cannot legitimately assert that the LSW answers are better than the CPS hot deck answers.

In the same vein, LSW's checking the accuracy of the LSW model by checking the prediction of respondents' values does not adequately check the imputations of the model for nonrespondents. In particular, both the ignorable and nonignorable nonresponse models discussed above will accurately reproduce the observed data for respondents, even though they predict very different amounts for nonrespondents. In order to address which model is more appropriate, we need data from nonrespondents or some external information about the distribution of reported incomes in the entire population.

9.7 The CPS-SSA-IRS Exact Match File

There is a data set that provides data relevant to accessing the differences in distributions of incomes between CPS nonrespondents and respondents. This data set is the CPS-SSA-IRS (SSA = Social Security Administration; IRS = Internal Revenue Service) Exact Match File (Aziz, Kilss, and Scheuren 1978). The exact match file is based on a sample of 1978 CPS interviews with incomes obtained from SSA and IRS administration records. Thus, this file is a data set consisting of CPS respondents and nonrespondents with administrative income always observed. By treating CPS nonrespondents' administrative income as missing and applying specific methods for handling nonresponse, we do in fact obtain some evidence for the adequacy of these specific techniques for adjusting for nonresponse bias, although admittedly for administrative income rather than CPS reported income. Both HR and GRZ compare results of their imputations to the administrative data for nonrespondents from the exact match file.

HR compares the imputations for social security benefits from a version of the CPS hot deck and those from an explicit two-stage log-linear/linear model and also evaluates the utility of multiple imputation for obtaining proper inferences. Since HR's objective is to predict social security benefits rather than total income, its results do not address the same kind of income nonresponse as studied in LSW.

GRZ, however, like LSW, studies earned income using maximum likelihood on essentially the same selection model as LSW with the restriction $\theta = 0$ (i.e., income is lognormal) and compares these predictions of nonrespondents' administrative income to their actual administrative income. Interesting conclusions of GRZ include: (a) the model

predicts nonrespondent income rather well; (b) the true residuals in the log scale for the entire population, although not normal, are approximately symmetric; and (c) the CPS hot deck underestimates income by about 7 percent. These results lend modest, although mixed, support to the utility of LSW/GRZ-type selection models for CPS income data.

The results of combining the efforts of LSW and GRZ by applying the extended LSW selection model to the exact match file would certainly be of interest. Of particular importance, such an application would help investigate which model for nonresponse is truly appropriate for CPS income data. Any such study would ideally include the use of multiple imputation so that variability can be properly assessed.

References

Aziz, F., B. Kilss, and F. Scheuren. 1978. *1973 current population survey—Administrative record exact match file codebook, part I, Code counts and item definitions*. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. *Journal Royal Statistical Society B* 26: 211–52.

Cochran, W. G., and D. B. Rubin. 1973. Controlling bias in observational studies: A review. *Sankhya – A, 35*, 4: 417–46.

Greenlees, J. S., W. S. Reece, and K. D. Zieschang. 1982. Imputation of missing values when the probability of response depends upon the variable being imputed. *Journal of the American Statistical Association* 77: 251–61.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–61.

Herzog, T. N., and D. B. Rubin. 1983. Using multiple imputations to handle nonresponse in sample surveys. In *Incomplete data and sample surveys*, vol. 2, *Theory and bibliographics*. D. B. Rubin, W. G. Medow, and I. Olkin (eds.), pp. 209–45. New York: Academic Press.

Lillard, L., J. P. Smith, and F. Welch. 1981. What do we really know about wages: The importance of non-reporting and census imputation. University of California at Los Angeles, unpublished paper.

Little, R. J. A. 1982. Models for response in sample surveys. *Journal of the American Statistical Association* 77: 237–50.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in the analysis of observational studies for causal effects. *Biometrika* 70, no. 1: 41–55.

Rubin, D. B. 1976*a*. Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics* 32, no. 1: 109–20. Printer's correction note p. 955.

———. 1976*b*. Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. *Biometrics* 32, no. 1: 121–32. Printer's correction note p. 955.

———. 1976*c*. Inference and missing data. *Biometrika* 63: no. 3: 581–92.

———. 1978*a*. Multiple imputations in sample surveys—A phenomenological Bayesian approach to nonresponse. With discussion and reply. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34. Also in *Imputation and editing of faulty or missing survey data*. U.S. Department of Commerce, 1–23.

———. 1978*b*. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* 7, no. 1: 34–58.

———. 1980*a*. Bias reduction using Mahalanobis' metric matching. *Biometrics* 36, no. 2: 295–98. Printer's correction p. 296.

———. 1980*b*. *Handling nonresponse in sample surveys by multiple imputations*. U.S. Department of Commerce, Bureau of the Census Monograph.

This Page Intentionally Left Blank